Inter Annotator Study

Choice of inter-annotator agreement measure:

Our annotation plan divides the data into four batches, with David and Daoming annotating a same subset from batches 1 and 2, Jacob and Nicole annotating a same subset from batches 3 and 4, and GPT-4o annotating the remaining data. To assess inter-annotator agreement, we perform pairwise comparisons both among human annotators and between human annotators and GPT-4o. Upon reviewing the annotations, we observed a significant class imbalance among the three labels. Specifically, most reviews were classified as "Pathos," while "Ethos" and "Logos" were assigned far less frequently. Given the pairwise comparison approach and the imbalance in label distribution, Cohen's κ is the most appropriate metric for our study, as it accounts for agreement expected by chance. We therefore have chosen Cohen's κ as our inter-annotator agreement measure.

Annotator Agreement Result:

Our annotator agreement calculation steps and result are shown in this python notebook (Note: Due to Nicole being on leave this week, we have temporarily excluded her annotations and inter-annotator agreement calculations. We plan to incorporate her annotations and update the agreement scores once she is available. However, we believe our current results still provide meaningful insights for this study.)

Surprisingly, both human-human and human-GPT agreement scores were lower than we expected. To understand the cause of this discrepancy, we looked into the possible causes for the poor agreement scores and proposed three potential solutions to address this conundrum.

Possible Causes

To understand the discrepancies between human-human agreement and human-GPT agreement, we have uncovered the following key findings:

1. Patterns of Disagreement:

   Human annotators only disagreed between Ethos vs. Pathos or Logos vs. Pathos, but never between Logos and Ethos. GPT, however, showed occasions in disagreeing with humans across all three labels. It suggests that the distinction between Ethos and Logos is pretty well-defined in human understanding, but not so much with GPT.

2. Inconsistencies in Annotations:

   Neither GPT nor human annotations were entirely consistent. Given the stochastic nature of Large Language Models (LLMs), it was unsurprising that GPT's labels varied across different runs. However, upon re-annotating, we found that even human annotators sometimes assigned different labels to the same review on different attempts.

Based on these findings, we identified two primary reasons for the discrepancies:

1. GPT Relies on Keywords Rather Than Context

GPT appears to classify reviews based on keyword detection rather than fully considering context and sentiment. When a sentence contains distinct entity names (e.g., places, movies, directors), the model is more likely to label it as "Ethos", regardless of actual intent.

Example: "Today I won free tickets to watch any movie at the theatre and I thought, why not watch Oppenheimer for the fourth time?" This review does not align with our Ethos definition, yet GPT labeled it as Ethos, likely due to the mention of "Oppenheimer".

2. Misalignment in Human Understanding of Labels

While GPT's inconsistency was expected, we were also concerned by the disagreements among human annotators and inconsistencies in our own labels. We suspect that this is due to a lack of shared, concrete definitions of Ethos, Pathos, and Logos—particularly in distinguishing Pathos and Logos/Ethos. Another important factor that may have led to this discrepancies in understanding is the nature of the movie reviews, that many of the reviews blurs the boundaries between the categories, or sometimes irrelevant to the movie at all (i.e. memes or jokes).

Possible Solution

Upon analyzing the possible causes behind the inconsistencies, we propose the following three potential solutions:

1. Incorporating One-Shot Examples from the Dataset

Currently, our prompt provides clear definitions and ideal examples for each label, but the model may struggle to generalize from these "perfect" examples to the more informal or "meme" reviews in our dataset. By including one-shot examples from our actual dataset—especially those that are more ambiguous—we may help the model better align with our intended labeling strategy.

2. Expanding the Labeling Dimensions

Some reviews do not clearly fit into Ethos, Pathos, or Logos, particularly humorous or meme-based content. Example: "Quirked-up physicist with a little bit of security clearance busts it down atomic style… is he goated with the Strauss?" GPT often mislabels such reviews as Logos or Ethos, likely due to the presence of references and structured language.

Introducing an additional category (e.g., "Humor/Meme") could provide an alternative label for these outliers, potentially improving the classification accuracy by both GPT and Human.

3. Exploring Human Annotation Feasibility

Since, as illustrated above, we have seen that human annotation is more effective in distinguishing Logos vs Ethos than GPT, if automated annotation remains unreliable, human annotation from Mechanical Turk could be a viable alternative.

However, while human annotation remains an option, we have no indication from our members' annotation that human annotation would

increase the agreement score. Thus, we have yet to assess the time and cost required to scale it effectively.